

On The Marriage of SPARQL and Keywords

Peng Peng, Lei Zou, Dongyan Zhao

Peking University, China;

{ pku09pp, zoulei, zhaodongyan}@pku.edu.cn

ABSTRACT

Although SPARQL has been the predominant query language over RDF graphs, some query intentions cannot be well captured by only using SPARQL syntax. On the other hand, the keyword search enjoys widespread usage because of its intuitive way of specifying information needs but suffers from the problem of low precision. To maximize the advantages of both SPARQL and keyword search, we introduce a novel paradigm that combines both of them and propose a hybrid query (called an SK query) that integrates SPARQL and keyword search. In order to answer SK queries efficiently, a structural index is devised, based on a novel integrated query algorithm is proposed. We evaluate our method in large real RDF graphs and experiments demonstrate both effectiveness and efficiency of our method.

1. INTRODUCTION

As more and more knowledge bases become available, the question of how end users can access this body of knowledge becomes of crucial importance. As the de facto standard of a knowledge base, RDF (Resource Description Framework) repository is a collection of triples, denoted as $\langle \text{subject}, \text{predicate}, \text{object} \rangle$. An RDF repository can be represented as a graph, where subjects and objects are vertices connected by labeled edges (i.e., predicates). Figure 1 shows an example RDF dataset and the corresponding RDF graph, which is a part of a well-known knowledge base Yago [27]. All subjects and objects correspond to vertices and predicates correspond to edge labels. The numbers besides the vertices are IDs, and they are introduced for the easy presentation.

As we know, SPARQL query language is a standard way to access RDF data and is based on the subgraph (homomorphism) match semantic [25]. Figure 2(a) shows an example of SPARQL query and its corresponding query graph is shown in Figure 1(c). The query semantics of the example SPARQL is “finding all actors starring in film *Philadelphia*”. To enable to use SPARQL, users should have full knowledge of the whole RDF schema. For example, users should know that predicate “actedIn” means “starring in” and the Philadelphia film’s URI is “Philadelphia(film)”. In real applications, it may not be practical to have full knowledge about the

whole schema; thus, it may not be possible to specify exact query criteria. The following example illustrates the challenges.

EXAMPLE 1. Find all actors starring in film *Philadelphia*, who are related to “Academy Award” and “Golden Globe Award”. Assume that we do not know the exact URIs corresponding to “Academy Award” and “Golden Globe Award”. Furthermore, there is no precise predicate corresponding to “related to”.

<p><i>SPARQL:</i></p> <pre>Select ?a where{ ?a type Actor. ?a actedIn Philadelphia(film). Philadelphia(film) type Film }</pre>	<pre>Select ?a where{ ?a ?p ?str. ?a type Actor. ?a actedIn Philadelphia(film). Philadelphia(film) type Film. FILTER regex(?str, "(Academy Award) Philadelphia(film) type Film} (Golden Globe Award)");</pre>
(a) Q_1	(b) Q_2

Figure 2: Example SPARQL Queries

There are two issues in this example. First, since we do not know the URIs of “Academy Award” and “Golden Globe Award”, we should provide a keyword search paradigm that maps the keywords to the corresponding entities or classes in RDF graphs. Existing SPARQL syntax only supports the regular expression, as shown in Figure 2. More typographic or linguistic distances, such as string edit distance [9] and google similarity distance [6], are desirable.

The second issue is that there is no precise predicate corresponding to “related to”. A possible solution is to use “unknown” predicate (i.e., a variable at the predicate position), but, it only finds one-hop relations. Figure 2(b) shows a SPARQL query with unknown predicate and the regular expression FILTER. However, it fails to finding the multiple-hop relations, which may also be informative to users. For example, Antonio Banderas, an actor starring in Philadelphia film, whose wife won “Golden Globe Award”. This is also a possible interesting result to users, but, this is a two-hop relation between Antonio Banderas and “Golden Globe Award”.

In contrast, keyword search [2, 17, 15, 10, 19, 18] on graphs provides an intuitive way of specifying information needs. For example, we input two keywords “Joanne Woodward” and “Golden Globe Award” to discover unbounded relations (i.e., the paths in RDF graphs) between them. However, keyword search may return a larger number of non-informative search answers to users.

In fact, users’ query intentions cannot be well modeled using a single query type in many real-life applications. Hence, a hybrid search capability is desired. In this paper, we propose an integrated query formulation (called a SPARQL-Keyword query, shorted as SK query) and the solution framework by combining advantages of SPARQL and keyword search. Generally speaking, the results of SK query is the k SPARQL matches that are closet to all keywords in RDF graph G , where k is a parameter given by users. The formal definition of SK query is given in Definition 2.2.

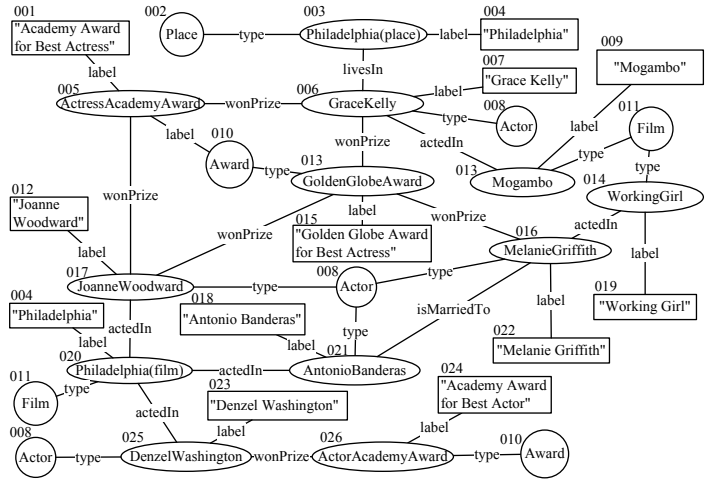
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WOODSTOCK '97 El Paso, Texas USA

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

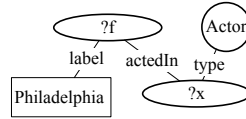
Subject	Predicate	Object
DenzelWashington	type	Actor
DenzelWashington	label	"Denzel Washington"
DenzelWashington	actedIn	Philadelphia(film)
DenzelWashington	wonPrize	AcademyAward
JoanneWoodward	type	Actor
JoanneWoodward	label	"Joanne Woodward"
JoanneWoodward	wonPrize	ActressAcademyAward
JoanneWoodward	wonPrize	GoldenGlobeAward
JoanneWoodward	actedIn	Philadelphia(film)
GraceKelly	type	Actor
GraceKelly	label	"Grace Kelly"
GraceKelly	actedIn	Mogambo
GraceKelly	wonPrize	ActressAcademyAward
GraceKelly	wonPrize	GoldenGlobeAward
GraceKelly	livesIn	Philadelphia(place)
AntonioBanderas	type	Actor
AntonioBanderas	label	"Antonio Banderas"
AntonioBanderas	actedIn	Philadelphia(film)
AntonioBanderas	isMarriedTo	MelanieGriffith
MelanieGriffith	type	Actor
MelanieGriffith	label	"William Holden"
MelanieGriffith	wonPrize	GoldenGlobeAward
MelanieGriffith	actedIn	WorkingGirl
ActressAcademyAward	type	Award
ActressAcademyAward	label	"Academy Award for Best Actress"
ActorAcademyAward	type	Award
ActorAcademyAward	label	"Academy Award for Best Actor"
GoldenGlobeAward	type	Award
GoldenGlobeAward	label	"Golden Globe Award for Best Actress"
Philadelphia(film)	type	Film
Philadelphia(film)	label	"Philadelphia"
WorkingGirl	type	Film
WorkingGirl	label	"Working Girl"
Mogambo	type	Film
Mogambo	label	"Mogambo"
Philadelphia(place)	type	Place
Philadelphia(place)	label	"Philadelphia"

(a) Example RDF Triples



(b) Example RDF Graph

SPARQL:



Keyword:

Academy Award
Golden Globe Award

(c) Example SPARQL-Keyword Query

Figure 1: RDF Examples

Let us recall Example 1 again. We issue the following SK query $\langle Q, q \rangle$. The SPARQL query graph Q is given in Figure 1(c), while the keywords are $q = \{\text{Academy Award, Golden Globe Award}\}$. Figure 3 shows three different results. First, there are three different subgraph matches of query Q , i.e., M_1 , M_2 and M_3 . Then, the keywords are matched in different literal vertices, i.e., 001, 015 and 026. The distance between a subgraph match M and a keyword in q is the shortest distance between M and one vertex containing keywords. We find that M_1 is the closest to the two keywords. It says “Joanne Woodward starring in Philadelphia film won both Academy Award and Golden Globe Award”. Obviously, this is an informative answer to the query in Example 1.

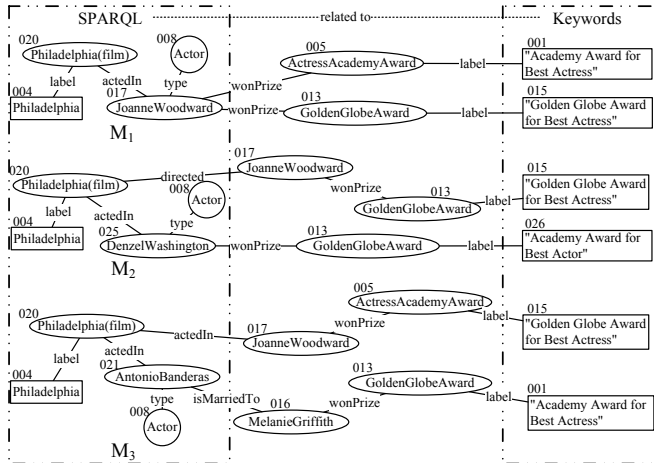


Figure 3: SK Query Results

In the above analysis, we assume that the relation strength depends on the path length, i.e., the number of hops. Actually, different predicates should have different weights to evaluate the relation strength. For example, there are two two-hops paths from

021 (AntonioBanderas) to 017 (JoanneWoodward). The first one is through 020 (Philadelphia(film)), while the second is through 008 (Actor). It is obvious that two people cooperating in the same film is more meaningful than both of them being actors, so the two-hops path through 020 has more relation strengths than the two-hops path through 008. Therefore, following the intuition of TF-IDF for measuring the word importance in a corpus, we propose “predicate salience” (see Section 2) to evaluate relation strengths.

Another challenge of this problem is the search efficiency. A naive “*exhaustive-computing*” strategy works as follows: we first find all subgraph matches of Q (in RDF graph G) by existing techniques; then, we compute the shortest path distances between these subgraph matches and the vertices containing keywords on the fly; finally, the matches with shortest distances to keywords are returned as answers. Obviously, this is an inefficient solution. Given a SPARQL query Q , there may exist some matches of Q that are far from the keywords in RDF graphs. These matches cannot contribute to the final results. Therefore, it is unnecessary to identify all subgraph matches in RDF graph. Instead of the exhaustive computing, we only find matches of SPARQL query Q progressively and design a lower bound that stops the search process as early as possible. Moreover, we propose a star index to enable the structural pruning.

In summary, we made the following contributions in this paper.

1. We propose a new query paradigm over RDF data combining keywords and SPARQL (called an SK query), and design a novel solution for this problem.
2. We design an index to speed up SK query processing. We propose a frequent star pattern-based index to reduce the search space.
3. We evaluate the effectiveness and efficiency of our method in real large RDF graphs and conclude that our methods are

much better than comparative models in both effectiveness (in terms of NDCG@k) and query response time.

The remainder of this paper is organized as follows: Section 2 defines the preliminary concepts. Section 3 gives an overview of our approach. We introduce a structural index to efficiently find the candidates of variables in SPARQL queries in Section 4. We discuss how to compute the results of SK queries in Section 5. Experimental results are presented in Section 6. Related works and the final conclusion are drawn in Section 7 and 8, respectively.

2. BACKGROUND

In this section, we introduce the fundamental definitions used in this paper.

2.1 Preliminaries

An RDF dataset consists of a number of triples, which is corresponding to an RDF graph. The SK query is to find k SPARQL matches that are top- k nearest with regard to all keywords.

DEFINITION 2.1. An RDF data graph G is denoted as $\langle V(G), E(G), L \rangle$, where (1) $V(G) = V_L \cup V_E \cup V_C$ is the set of vertices in RDF graph G , where V_L , V_E and V_C denote literal vertices, entity vertices and class vertices, respectively; (2) $E(G)$ is the set of edges in G ; and (3) L is a finite set of edge labels, i.e. predicates.

DEFINITION 2.2. A SK (SPARQL & Keyword) query is a pair $\langle Q, q \rangle$, where Q is a SPARQL query graph and q is a set of keywords $\{w_1, w_2, \dots, w_n\}$.

Given an SK query $\langle Q, q \rangle$, the result of $\langle Q, q \rangle$ in a data graph G is a pair $\langle M, \{v_1, v_2, \dots, v_n\} \rangle$, where M is a subgraph match of Q in G and v_i ($i = 1, \dots, n$) is a literal vertex (in G) containing keyword w_i .

Given an SK query $\langle Q, q = \{w_1, \dots, w_n\} \rangle$, the cost of a result $r = \langle M, \{v_1, v_2, \dots, v_n\} \rangle$ contains two parts. The first part is the content cost and the second part is the structure cost.

DEFINITION 2.3. Given a result $r = \langle M, \{v_1, v_2, \dots, v_n\} \rangle$, the cost of r is defined as follows:

$$Cost(r) = Cost_{content}(r) + Cost_{structure}(r)$$

where $Cost_{content}(r)$ is the content cost of r (defined in Definition 2.4) and $Cost_{structure}(r)$ is the structure cost of r (defined in Definition 2.5).

DEFINITION 2.4. Given a result $r = \langle M, \{v_1, v_2, \dots, v_n\} \rangle$, the content cost of $r = \langle M, \{v_1, v_2, \dots, v_n\} \rangle$ is defined as follows:

$$Cost_{content}(r) = \sum_{i=1}^{i=n} C(v_i, w_i)$$

where $C(v_i, w_i)$ is the matching cost between v_i and keyword w_i .

Any typographic or linguistic distances, such as string edit distances [31] and google similarity distance [7], can be used to measure $C(v_i, w_i)$.

In applications, users are more interested in some variables (in SPARQL query Q) than the constants in Q . Let us recall Example 1. The distance between keywords and the matching vertices with regard to variable “?a” is more interesting to measure the relationship strength. Therefore, to evaluate the structure cost (in Definition 2.5), we only consider the matching vertices with regard to variables in SPARQL query Q .

DEFINITION 2.5. Given a result $\langle M, \{v_1, v_2, \dots, v_n\} \rangle$ for an SK query $\langle Q, q \rangle$, the distance between match M and vertex v_i ($i = 1, \dots, n$) is defined as follows.

$$d(M, v_i) = \min_{v \in M} \{d(v, v_i)\}$$

where v is a matching vertex in M with regard to a variable in SPARQL query Q and $d(v, v_i)$ is the shortest path distance between v and v_i in RDF graph G .

Then, the structure cost of a result $r = \langle M, \{v_1, v_2, \dots, v_n\} \rangle$ is defined as follows.

$$Cost_{structure}(r) = \sum_{i=1}^{i=n} d(M, v_i)$$

(Problem Definition) Given an SK query $\langle Q, q \rangle$ and a parameter k , our problem is to find k results (Definition 2.2), which have the k -smallest costs.

2.2 Predicate Saliency

In this paper, we use “shortest path distance” to evaluate the relation strength. However, the naive definition of the shortest path distance suffers from a critical problem: all predicates, i.e., edge labels, are considered equally important when it is used to measure the relationship strength between entities. In fact some predicates have little or no discriminating power in determining relevance. For example, predicates like “type” and “label” are so common that each entity is incident to a class vertex through an edge of predicate “type”. This tends to incorrectly emphasize paths which contain these common predicates more frequently, without giving enough weight to the paths of more meaningful predicates (like “actedIn” and “isMarriedTo”). The predicates like “type” and “label” are not good predicates to distinguish relevant and non-relevant vertices, unlike the less common predicates “actedIn” and “isMarriedTo”.

Hence, we should introduce a mechanism for attenuating the effect of predicates that occur too frequently in the RDF graph to be meaningful for relevance determination. Learning from the concept of document frequency, we first find out the set of vertices in the RDF graph incident to a predicate p , which is denoted as $V(p)$. Then, we divide the size of $V(p)$ by the total number of vertices. We name the measure as the *predicate saliency* of predicate p and give the formal definition as follows:

$$ps(p) = \frac{|V(p)|}{|V(G)|}$$

Thus the predicate saliency of a rare predicate is low, whereas the predicate saliency of a frequent predicate is likely to be high, which means that rare predicates have less cost than frequent predicates.

Let us consider RDF graph in Figure 1(b). The predicate saliency values of all predicates are given in Table 1. As shown in Table 1, predicate “actedIn” is more important than “type” in measuring the relation strength, while the former’s predicate saliency is 0.296 and the latter’s is 0.593.

Predicate	Predicate Saliency
actedIn	0.296
isMarriedTo	0.074
label	0.852
livesIn	0.074
type	0.593
wonPrize	0.259

Table 1: Weights of Predicates in Example RDF Graph

3. OVERVIEW

In this section, we give an overview of the different steps involved in our process of SK query, which is depicted in Figure

4. In this paper, we are concerned with the challenge of efficiently finding the results of SK queries. We propose an approach in which the best results of the SK query are computed using the graph exploration. We detail the different steps of the approach below.

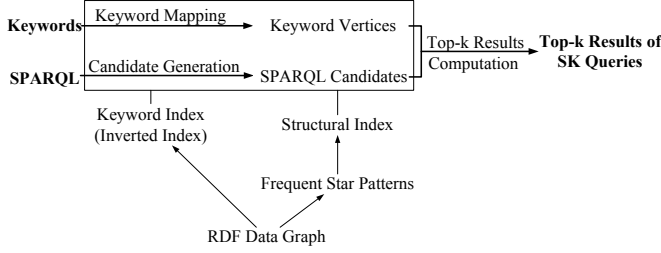


Figure 4: Overview of Our Approach

Keyword Mapping. In the offline phase, we create an inverted index storing a map from keywords to its locations in the RDF graph. In the online phase, we map keywords to vertices based on the inverted index.

For scoring keyword vertices, a widely used metric that is computed on-the-fly for a given query is IR-style TF/IDF cost. Many cost functions have been proposed in the literature, and we select one of them to assign the cost to each vertex containing keywords. Note that we need to normalize the cost of keyword matching vertices before the distances computation.

In this paper, our primary focus is indexing and query processing, so we will not delve into the specifics of keyword mapping.

Candidate Generation. When we find a vertex reachable to elements of all keywords, we need to run subgraph homomorphism to check whether there exist some subgraph matches (of Q) containing v . As we know, subgraph homomorphism is not efficient due to its high complexity [14]. In order to speed up query processing, we propose a filter-and-refine strategy to reduce the number of subgraph homomorphism operations. The basic idea is to filter out some vertices that are not in any subgraph match of Q . We call them “dummy” vertices. If the search meets a dummy vertex, we do not perform subgraph homomorphism algorithm.

In this paper, we propose a frequent star pattern-based structural index. Based on this index, we can locate a candidate list in RDF graph of each variable in SPARQL. A vertex in at least one candidate lists of variables is not dummy. We will detail how to build the structural index in Section 4.1 and how to use the index to reduce the candidates of all variables in Section 4.2.

Top-k Results Computation. Based on the keyword vertices and variables’ candidates, we propose a solution based on graph exploration to compute the top-k result of SK queries. Our approach starts graph exploration from all keyword vertices, and explores to their neighboring vertices recursively until the distances between a vertex and keyword vertices have been computed out. When the distances between a vertex and vertices of all keywords have been computed out, we check whether this vertex is a dummy vertex. If so, there exists no match of Q that can contain it. Hence, we can skip it. Otherwise, we start our SPARQL matching algorithm (Algorithm 2) from the vertex to generate all matches containing it. The exploration terminates when the top-k results have been computed. We also propose some early stop strategies for top-k computation to reach early termination after obtaining the top-k results, instead of searching the data graph for all results.

We discuss the detail of top-k results computation in Section 5.

4. CANDIDATE GENERATION BASED ON STRUCTURAL INDEX

In this section, we first introduce an structural index based on a certain kind of patterns in Section 4.1. Then, we discuss how to generate the candidate lists of variables based on our structural index 4.2.

4.1 Structural Index

In this section, we propose a frequent star pattern-based index. We mine some frequent star patterns in G . For each frequent star S , we build an inverted list $L(S)$ that includes all vertices (in RDF graph G) contained by at least one match of S . A reason for selecting stars as index elements is that SPARQL queries tend to contain star-shaped subqueries, for combining several attribute-like properties of the same entity [22].

We propose a sequential pattern mining-based method to find frequent star patterns in RDF graphs. For each entity vertex in an RDF graph, we sort all its adjacent edges in lexicographic order of edge labels (i.e. properties). These sorted edges can form a sequence. For example, vertex “Philadelphia(film)” has five adjacent edges, that are $\langle actedIn, actedIn, actedIn, name, type \rangle$. Table 2 shows a sequence database, where each sequence is formed by the adjacent edges of one entity vertex. We employ the existing sequential pattern mining algorithms, such as PrefixSpan[24], to find frequent sequential patterns, where each sequential pattern corresponds to a star pattern in RDF graphs. For example, assume that the minimal support count $s = 2$, $\langle actedIn, type \rangle$ and $\langle actedIn, type, wonPrize \rangle$ are two frequent sequential patterns. It is easy to know that a sequential pattern always corresponds to one star pattern, as shown in Figure 5. For ease of presentation, we use the terms “sequential patterns” and “star patterns” interchangeably in the following discussion.

Vertex	Predicate Sequence
Philadelphia(film)	$\langle actedIn, actedIn, actedIn, label, type \rangle$
JoanneWoodward	$\langle actedIn, actedIn, label, type, wonPrize \rangle$
AntonioBanderas	$\langle actedIn, label, type, wonPrize, wonPrize \rangle$
...	...

Table 2: Example of Predicate Sequences

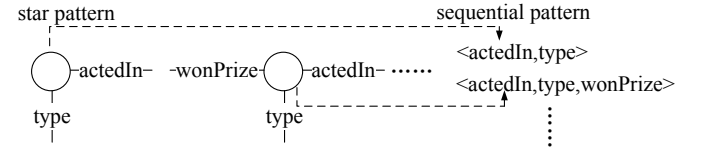


Figure 5: Example Star Pattern

For each frequent star pattern S , we maintain an inverted list $L(S) = \{v | S \text{ occurs in } v\text{'s adjacent edge sequence}\}$. Obviously, if we use all frequent stars as the index elements, the space cost is very large. Thus, inspired by gIndex [37], we also define the *discriminative ratio* for the star pattern selection.

DEFINITION 4.1. Given a star S , its discriminative ratio is defined as follows:

$$\gamma(S) = \frac{|L(S)|}{|\bigcap_{S' \subset S} L(S')|}$$

where $S' \subset S$ denotes that S' is a part of S .

Obviously, $\gamma(S) \leq 1$. $\gamma(S) = 1$ means that $L(S)$ can be obtained by the intersection of all $L(S')$, where $S' \subset S$. In this case, if all S' are index elements, it is not necessary to keep S as the index element, as S cannot provide more pruning power. In practice, we set a threshold γ_{max} , and we only choose the stars S , where $\gamma(S) \leq \gamma_{max}$. Note that, to ensure the completeness of the indexing, we always choose the (absolute) support to be 1 for size-1 stars (star with only one edge). This method can guarantee no-false-negative, since all vertices (in G) are indexed in at least one inverted list.

THEOREM 4.1. Let F denote all selected index elements (i.e., frequent star patterns). Given a SPARQL query Q , a vertex v in graph G can be pruned (there exists no subgraph match of Q containing v) if the following equation holds.

$$v \notin \bigcup_{S \in F \wedge S \in Q} L(S) \quad (1)$$

where $S \in F$ means that S is a selected star pattern and $S \in Q$ is a star pattern included in Q .

PROOF. If $v \notin \bigcup_{S \in F \wedge S \in Q} L(S)$, it means that the structure around v does not contain any substructure of Q . Hence, v must be unable to in a match of Q . \square

4.2 Candidate Generation

Given an SK query, we first tag the vertices that can be pruned by Theorem 4.1. For each variable in SPARQL, we locate its candidates in RDF graph. Each variable can map to a predicate sequence according to the SPARQL statement. For example, variable “?a” of the SPARQL query in Figure 1(c) has the predicate sequence $\langle \text{actedIn}, \text{type} \rangle$. Then, for each variable x , we look up our structural index and find the maximum pattern contained by x ’s predicate sequence. We load the vertex list of the maximum pattern as x ’s candidates. A vertex in at least one vertex lists of variables is not dummy. We define these pruned vertices as *dummy vertices* as follows.

DEFINITION 4.2. Dummy Vertex. Given a SPARQL query Q , a vertex v in graph G is called as dummy vertex if the following equation holds.

$$v \notin \bigcup_{S \in F \wedge S \in Q} L(S) \quad (2)$$

where F denote all selected frequent star patterns, $S \in F$ means that S is a selected star pattern and $S \in Q$ is a star pattern included in Q .

When the search process meets a fully-seen vertex v , if v is not a dummy vertex, we perform subgraph isomorphism algorithm to find the subgraph match of SPARQL query Q containing v . Otherwise, we do not perform subgraph isomorphism algorithm beginning with v .

5. TOP-K RESULTS COMPUTATION

In this section, we introduce our approach for SK queries, which is based on the “backward search” strategy [2]. Our algorithm for searching top-k results of SK queries is shown in Algorithm 1. This algorithm consists of three parts: 1) graph exploration to find vertices connecting the keyword vertices, 2) generation of SPARQL matches from the vertices connecting the keyword vertices and 3) top-k computation. In the following, we will elaborate on these three tasks.

5.1 Graph Exploration

Given the keyword vertices, the objective of the exploration is to find vertices in the graph that connect these keyword vertices and compute their distances to these keyword vertices. Let V_i denote all literal vertices (in RDF graph G) containing keyword w_i .

DEFINITION 5.1. Distance between Vertex and Keyword. Given a vertex v in RDF graph G and a keyword w_i , the distance between v and keyword w_i (denoted as $d(v, w_i)$) is the minimum distance between v and a vertex in V_i , where V_i includes all literal vertices containing keyword w_i in G .

For graph exploration, we maintain a priority queue PQ_i for each keyword w_i . Each element in PQ_i is represented as $(v, p, |p|)$, where v is a vertex id, p is a path between v and a vertex in V_i and $|p|$

Algorithm 1: Search for Top-k Results of SK Queries

Input: RDF data graph G , SK query $\langle Q, q \rangle$, $\{V_1, \dots, V_n\}$ where V_i is the set of vertices containing keyword w_i , priority queues $\{PQ_1, \dots, PQ_n\}$.

Output: Top-k results R of $\langle Q, q \rangle$.

```

1 for each vertices set  $V_i$  do
2   for each vertex  $v$  in  $V_i$  do
3     Insert  $(v, \emptyset, 0)$  into  $PQ_i$ ;
4 while not all queues are empty do
5   for  $i = 1, \dots, n$  do
6     Pop the head of  $PQ_i$  (denoted as  $(v, p, |p|)$ ), set  $d[v][i] = |p|$ 
       and insert it into  $RS_i$ ;
7     for each adjacent edge  $vv'$  to  $v$  do
8       if  $p \cup vv'$  is not a simple path then
9         Continue;
10      if there exists another element  $(v', p', |p'|)$  in  $PQ_i$  then
11        if  $|p'| > |p| + ps(vv')$  then
12          Delete  $(v', p', |p'|)$ ;
13          Insert  $(v', p \cup vv', |p| + ps(vv'))$  in  $PQ_i$ ;
14        else
15          Continue;
16      else
17        Insert  $(v', p \cup vv', |p| + ps(vv'))$  in  $PQ_i$ ;
18      if  $v$  is a fully-seen vertex then
19        Call Algorithm 2 to find all matches containing  $v$ ;
20      for each match  $M$  containing vertex  $v$  do
21        if all vertices in  $M$  are fully-seen vertices then
22          Use  $M$  to update  $R$  and the upper bound  $\delta$  of top-k
           results
23        Update the cost of all partially-seen matches and  $\delta$ ;
24        Update the lower bound cost  $\theta$  of all remaining un-seen
           vertices;
25      if  $\theta \geq \delta$  then
26        Break;
27 Return  $R$ .
```

denotes the path distance. All elements in PQ_i are sorted in the non-descending order of $|p|$. Each keyword w_i is also associated with a result set RS_i . In order to keep track of information related to each vertex v , we associate v with a vector $d[v]$. If a vertex v is in RS_i ($i = 1, \dots, n$), the shortest path distance is known. In this case, we set $d[v][i] = d(v, w_i)$; otherwise, we set $d[v][i] = \text{null}$.

Initially, the exploration starts with a set of vertices containing keywords. For each vertex v containing keyword w_i , an element $(v, \emptyset, 0)$ is created and placed into the queue PQ_i (Line 3 in Algorithm 1). During the search, at each step, we pick a queue PQ_i ($i = 1, \dots, n$) to expand in a round-robin manner (Line 5 in Algorithm 1). We assume that we pop the queue head $(v, p, |p|)$ from PQ_i . When a queue head $(v, p, |p|)$ is popped from queue PQ_i , we insert it into result set RS_i and set $d[v][w_i] = |p|$ (Line 6 in Algorithm 1). We can prove that the following theorem holds.

THEOREM 5.1. When a queue head $(v, p, |p|)$ is popped from queue PQ_i , the following equation holds.

$$d(v, w_i) = d[v][i] = |p|$$

PROOF. Given a vertex v , before $(v, p, |p|)$ is popped from PQ_i , a path p between v and vertices containing w_i . It is obvious that $|p| \geq d(v, w_i)$.

We wish to show that in each iteration, $d(v, w_i) = d[v][i] = |p|$ for the element $(v, p, |p|)$ popped from PQ_i . We prove this by contradiction. We assume that v is the first vertex for which $d[v][i] = |p| \neq d(w_i, v)$ when $(v, p, |p|)$ is popped from PQ_i . We focus our attention on the situation at the beginning of the iteration in which $(v, p, |p|)$ is popped from PQ_i and derive the contradiction that $d[v][i] = |p| = d(v, w_i)$ at that time by examining the shortest path from v to ver-

tices containing w_i . We must have $v \notin V_i$ because all vertices in V_i are the first vertices added to set RS_i and $d[v][i] = 0$ at that time.

Because $v \notin V_i$, we also have that $RS_i \neq \emptyset$ just before $(v, p, |p|)$ is popped from PQ_i . There must be some paths from vertices containing w_i to v , for otherwise $d[v][i] = \infty$ by the no-path property, which would violate our assumption that $d[v][i] \neq d(w_i, v)$. Because there is at least one path, there is the shortest path p' between v and vertices in V_i . Prior to pop $(v, p, |p|)$ to PQ_i , path p' connects a vertex in RS_i , namely some vertices in V_i , to a vertex in $V(G) - RS_i$, namely v . Let us consider the first vertex v' along p' such that $v' \in V(G) - RS_i$, and let $v'' \in RS_i$ be the predecessor of v' .

We claim that $d[v'][i] = d(w_i, v')$ when the element of v' is popped from PQ_i . To prove this claim, observe that $v'' \in RS_i$. Then, because v is chosen as the first vertex for which $d[v][i] \neq d(w_i, v)$ when $(v, p, |p|)$ is popped from PQ_i , we have $d[v'][i] = d(w_i, v')$ when v' is added to RS_i . Edge $v''v'$ is relaxed at that time (Line 7 - 17 in Algorithm 1), so the claim follows from the convergence property.

We can now obtain a contradiction to prove that $d[v][i] = d(v, w_i)$. Because v' occurs before v on the shortest path from vertices in V_i to v and all edge weights are nonnegative, we have $d[v'][i] \leq d(v, w_i)$, and thus $d(v', w_i) = d[v'][i] \leq d(v, w_i) \leq d[v][i]$.

But because both vertices v and v' are in $V(G) - RS_i$ when v' is popped before v , we have $d(v', w_i) \leq d(v, w_i)$. Thus, $d(v', w_i) = d[v'][i] = d(v, w_i)$, which contradicts our choice of v . We conclude that $d[v][i] = d(w_i, v)$ when $(v, p, |p|)$ is popped from PQ_i , and that this equality is maintained at all times thereafter. \square

When a queue head $(v, p, |p|)$ is popped from queue PQ_i , it means that we have computed out the distance between v and keyword w_i . We also say that v is *seen* by keyword w_i .

DEFINITION 5.2. Seen by Keyword. When a queue head $(v, p, |p|)$ is popped from queue PQ_i , we say vertex v is seen by keyword w_i .

Assume that $(v, p, |p|)$ is popped from queue PQ_i . For each incident edge $\overline{vv'}$ to v , we obtain a new element $(v', p \cup \overline{vv'}, |p| + ps(\overline{vv'}))$, where $p \cup \overline{vv'}$ means appending an edge to p and $ps(\overline{vv'})$ denotes the predicate salience value of the edge label of $\overline{vv'}$. If $p_i \cup \overline{vv'}$ is not a simple path¹, the element is ignored (Line 8-9 in Algorithm 1). Then, we check whether there exists another element $(v', p', |p'|)$ that has the identical vertex v' with the new element $(v', p \cup \overline{vv'}, |p| + ps(\overline{vv'}))$, where $|p'| > |p| + ps(\overline{vv'})$. If so, we delete $(v', p', |p'|)$ from PQ_i and insert $(v', p \cup \overline{vv'}, |p| + ps(\overline{vv'}))$ into PQ_i (Line 12-13 in Algorithm 1). Otherwise, we ignore the new element (Line 15 in Algorithm 1). If there exists no element $(v, p', |p'|)$, we insert $(v', p \cup \overline{vv'}, |p| + ps(\overline{vv'}))$ into the queue directly (Line 17 in Algorithm 1).

DEFINITION 5.3. Fully-seen Vertex, Partially-seen Vertex and Un-seen Vertex. Given a vertex v , if v is seen by all keywords w_i ($i = 1, \dots, n$), v is called a *fully-seen vertex*; if v is not a *fully-seen vertex* but it has been seen by at least one keyword, v is called a *partially-seen vertex*; if v is not seen by any keyword, v is called an *un-seen vertex*.

At each step, we check whether the vertex just popped from the queue has been seen by all keywords. Specifically, for a popped queue head v , if all dimensions of its vector $d[v]$ are non-null, it means that all keywords have seen vertex v , i.e., we have known the distance between v and each keyword. In this case, v is a *fully-seen vertex*. When we meet a fully-seen vertex v , we will employ a subgraph homomorphism algorithm to find matches containing v (Line 18-19 in Algorithm 1). The details will be discussed in Section 5.2.

¹A simple path is a path with no repeated vertices.

Algorithm 2: SPARQL Matching Algorithm

Input: A candidate vertex v corresponding to u in SPARQL query Q , and a state stack S .
Output: The match set MS of Q containing v .

```

1 Initialize a state  $s$  with  $v$ ;
2 Push  $s$  into  $S$ ;
3 while  $S \neq \emptyset$  do
4   Pop the first state  $s \in S$ ;
5   if all edges of  $Q$  have been matched in  $s$  then
6     Insert  $s$  to  $MS$ ;
7   for each unmatched edge  $\overline{u'u''}$  that  $u'$  has been matched to  $v'$  do
8     if  $u''$  has been matched to  $v''$  then
9       if  $v''v''' \in E(G)$  then
10        Push  $s$  into  $S$ ;
11      else
12        Continue;
13    else
14      for each neighbor  $v''$  of  $v'$  do
15        if  $v''$  is a dummy or fully-seen vertex then
16          Continue;
17        if  $v''v'''$  can match  $\overline{u'u''}$  then
18          Initialize a new state  $s'$  and  $s' = s$ ;
19          Match  $u''$  with  $v''$ ;
20          Push  $s$  into  $S$ ;
21 Return  $MS$ .
```

5.2 Generation of SPARQL Matches

When we find out a fully-seen vertex v , it means that we have known the distance between v and each keyword w_i . The next step is to compute SPARQL matches containing vertex v if any. Here, we perform subgraph homomorphism algorithm to find subgraph matches (of query Q) containing v .

Generally speaking, we employ a DFS-based *state transformation* algorithm to perform the matching process beginning from a fully-seen vertex v (as shown in Algorithm 2). Here, we define the *state* as follows.

DEFINITION 5.4. Given a SPARQL query graph Q with m vertices u_1, \dots, u_m , a state is a (partial) match of query graph Q .

For example, Figure 6 shows an example state of the SPARQL query in Figure 1(c).

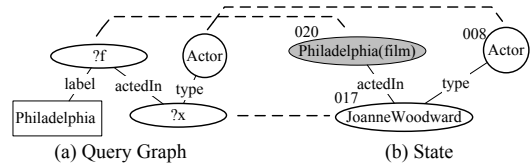


Figure 6: A State Matching a Part of Q

In particular, our *state transformation* algorithm is as follows. Assume that v matches vertex u in SPARQL query Q . We first initialize a state with v . Then, we search the RDF data graph to reach v 's neighbor v' corresponding to u' in Q , where u' is one of u 's neighbors and edge $\overline{vv'}$ satisfies query edge $\overline{uu'}$. The search will extend the state step by step. The search branch terminates until that we have found a state corresponding to a match or we cannot continue. In this case, the algorithm is backtrack to some other states and try other search branches.

As shown in Algorithm 2, we find all matches containing some fully-seen vertex v only if v is not be a dummy vertex (Line 15 - 16 in Algorithm 2). This is because that there exists no subgraph match containing a dummy vertex. When we finish Algorithm 2 from v , we say that v is *searched*. The "searched" indicates that

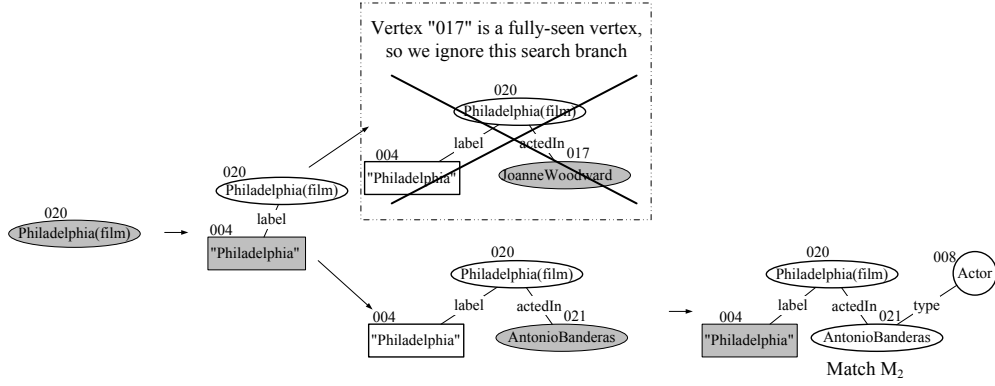


Figure 7: Finding Matches Containing Vertex “020” by Pruning the Search Branch Beginning from “017”

all matches containing v has been found if any. When we search the RDF graph beginning with a fully-seen vertex, if the search meets another fully-seen vertex v'' , it can skip v'' (Line 15 - 16 in Algorithm 2). This is because that the matches containing v'' have been found before.

EXAMPLE 2. We assume that the current popped fully-seen vertex is “020” and vertex “017” is another a fully-seen vertex. As shown in Figure 7, we explore the RDF graph from “020” to “017”. However, vertex “017” is a fully-seen vertex, so all SPARQL matches containing “017” have been found already. As a result, we can terminate the corresponding search branches in Figure 7.

5.3 Top-k Computation

The native solution for computing top-k results of a SK query is to run the backward search algorithm until that all vertices (in RDF graph G) have been fully-seen by keywords. Then, according to the results’ cost, we can find the top-k results. Obviously, this is an inefficient solution especially when G is very large. In this subsection, we design an early-stop strategy.

Let us consider a snapshot of some iteration step in Algorithm 1. All subgraph matches of SPARQL query Q can be divided into three categories: fully-seen matches, partially-seen matches and un-seen matches.

DEFINITION 5.5. Fully-seen Match, Partially-seen Match and Un-seen Match. Given a subgraph match M of SPARQL query Q , if all vertices in M are fully-seen vertices, M is called a fully-seen match; if M is not a fully-seen match and M contains at least one fully-seen vertex, it is called a partially-seen match. If a match M does not contain any fully-seen vertex, it is called an un-seen match.

Figure 8 demonstrates a visual representation of three kinds of matches. The shaded area covered by the dash line circle denotes all fully-seen vertices in RDF graph. With the increasing of the iteration steps (in Algorithm 1), the shaded area expands gradually until that it covers the whole RDF graph. The early-stop strategy is to stop the expansion as early as possible, but we can guarantee that we have found the top-k results for SK queries.

The basic idea of our early-stop strategy is as follows. We only compute the cost of fully-seen matches. Then, we use the fully-seen matches to find a threshold δ , which is the k -th smallest cost so far. If there are less than k fully-seen matches so far, δ is ∞ . We compute the lower bounds θ_1 and θ_2 for partially-seen matches and un-seen matches, respectively. The algorithm can early stop if and only if $\delta < \theta_1 \wedge \delta < \theta_2$. Otherwise, the algorithm continues the next iteration.

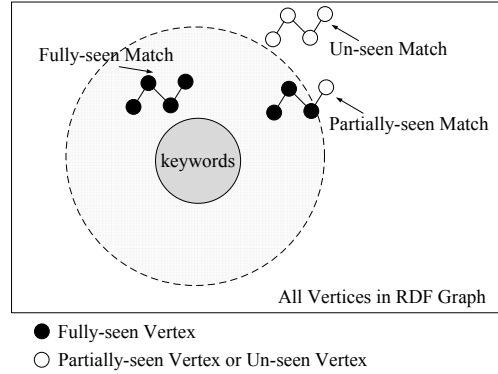


Figure 8: Fully-seen Match, Partially-seen Match and Un-seen Match during the Backward Search

Fully-seen Match. For a fully-seen match, we compute its match cost according to Definition 2.5. If we have found more than k fully-seen matches, we maintain a threshold δ be the k -th smallest match cost.

Partially-seen Match. For any partially-seen match, we compute the lower bound of its cost as follows.

THEOREM 5.2. Given a partially-seen match M of SPARQL query Q , v is a partially-seen vertex or an un-seen vertex in the match. The following equation holds.

$$\begin{aligned} \text{Cost}(M) &= \sum_{1 \leq i \leq n} d(v, w_i) \\ &\geq \sum_{d[v][w_i] \neq \text{null} \wedge 1 \leq i \leq n} d[v][w_i] + \sum_{d[v][w_i] = \text{null} \wedge 1 \leq i \leq n} |p_i| \end{aligned}$$

where $d[v][w_i]$ is the i -th dimension of v 's vector corresponding to keyword w_i , and $|p_i|$ corresponds to the current queue head $(v, p_i, |p_i|)$ in queue PQ_i .

PROOF. If $d[v][i] \neq \text{null}$, it means that we have computed out $d(v, w_i)$. If $d[v][w_i] = \text{null}$, v has still not been seen. Since each time we pop the head $(v, p_i, |p_i|)$ of PQ_i where $|p_i|$ is the smallest, all un-seen vertices' distances to w_i are larger than $|p_i|$. \square

According to Theorem 5.2, we define the lower bound of a partially-seen match M as follows.

DEFINITION 5.6. Given a match of SPARQL query Q , the lower bound for a partially-seen match M is defined as follows.

$$lb(M) = \min_{v \in M} \left(\sum_{d[v][w_i] \neq \text{null} \wedge 1 \leq i \leq n} d[v][w_i] + \sum_{d[v][w_i] = \text{null} \wedge 1 \leq i \leq n} |p_i| \right)$$

The lower bound for all partially-seen matches is defined as follows.

DEFINITION 5.7. *The lower bound θ_1 for all partially-seen matches is as follows.*

$$\theta_1 = \min_{M \in PS} (lb(M))$$

where PS denotes all partially-seen matches and $lb(M)$ is defined in Definition 5.6.

With the increasing of the iteration steps, some partially-seen matches become fully-seen matches. They are moved to FS . The threshold δ and θ_1 are updated accordingly.

Un-seen Match. Let us consider an un-seen match M . There are two kinds of vertices in M , i.e., partially-seen vertices and un-seen vertices.

THEOREM 5.3. *For an un-seen vertex v , if threshold $\delta \neq \infty$, the following equation holds.*

$$\delta \leq \sum_{1 \leq i \leq n} d(v, w_i)$$

PROOF. For each keyword w_i , we assume that the queue head of PQ_i is $(v, p_i, |p_i|)$. Since v is an un-seen vertex, $|p_i| \leq d(v, w_i)$ for each keyword w_i . In contrast, δ is the upper bound of the top-k results, so δ is equal to the cost of a fully-seen match M . Each vertex v' in M is fully-seen vertex. Hence, $d(v', w_i) \leq |p_i|$. Then, we know that $d(v', w_i) \leq |p_i| \leq d(v, w_i)$ for each keyword w_i . In conclusion, $\delta \leq \sum_{1 \leq i \leq n} d(v, w_i)$. \square

According to Theorem 5.3, it is not necessary to consider un-seen vertices to define the lower bound for un-seen matches. Therefore, we define the lower bound for all un-seen matches as follows.

DEFINITION 5.8. *The lower bound θ_2 for all un-seen matches is as follows.*

$$\theta_2 = \min_{v \in PSet} \left(\sum_{d[v][w_i] \neq null \wedge 1 \leq i \leq n} d[v][w_i] + \sum_{d[v][w_i] = null \wedge 1 \leq i \leq n} |p_i| \right)$$

where $PSet$ contains all partially-seen vertices so far, $d[v][w_i]$ is the i -th dimension of the v 's vector corresponding to keyword w_i and $|p_i|$ corresponds to the current queue head $(v, p_i, |p_i|)$ in queue PQ_i .

Early-stop Strategy. In each iteration step, we check whether $\delta \leq \theta_1 \wedge \delta \leq \theta_2$. If the condition holds, the algorithm can stop, since any partially-seen match or un-seen match cannot be in one of the top-k results.

6. EXPERIMENTS

In this section, we evaluate our approach in three large real RDF graphs, DBLP, Yago and DBPedia.

For effectiveness study, we compare our method with a classical keyword search algorithm BANKS [2] over both Yago and DBPedia. Furthermore, since each resource in DBPedia is annotated by Wikipedia documents, so we design a stronger baseline named as “Annotated SPARQL” for DBPedia. “Annotated SPARQL” is similar to the approach discussed in [33]. It first finds out all matches of the SPARQL query, then ranks these matches by how closely the corresponding Wikipedia documents match the keywords. Note that, except for DBPedia, most current RDF datasets do not provide such documents to annotate the resources. Hence, we only do experiments of annotated SPARQL over DBPedia. For other RDF datasets, although we can crawl some pages to annotate their entities, that is beyond the scope of this paper.

For efficiency study, because there are no existing method for SK queries, we evaluate our approach with two baselines, i.e., the *exhaustive computing* and the *naive backward search* and . “Exhaustive computing” has been introduced in Section 1. The *naive backward search* is to run the backward search algorithm until that all vertices have been fully-seen by keywords. Then, according to the results’ cost, we find the top-k results.

Our experiments are conducted on a machine with 2 Ghz Core 2 Duo processor, 16G RAM memory and running Windows Server 2008. All experiments are implemented in Java languages. We use MySQL to store the RDF graphs and the indices.

6.1 Datasets & Setup

We use three real-world RDF datasets, DBLP, Yago and DBPedia in our experiments. The details about the two datasets are as follows.

DBLP. DBLP² contains bibliographic information of computer science publications [21]. The DBLP graph contains 8,381,858 RDF triples and 3,103,614 vertices. We define 5 sample SK queries for DBLP and show two of them in Table 3 for case study.

Yago. Yago³ extracts facts from Wikipedia and integrates them with the WordNet thesaurus [27]. The RDF graph has 19,012,849 edges and 12,811,222 vertices. We define 8 sample SK queries for Yago and show two of them for case study in Table 4.

DBPedia & QALD. DBPedia⁴ is an RDF dataset extracted from Wikipedia. The DBPedia contains 73,766,900 edges and 13,100,739 vertices. QALD⁵ is an evaluation campaign on question answering over linked data. It is co-located with the ESWC 2012. In this campaign, the committee provides some questions and each question is annotated with some recommended keywords and the answers that these queries retrieve. Note that, some questions in QALD are so simple that they can map to a SPARQL query with only one edge. These simple questions are unnecessary to be split into a SPARQL query and some keywords. Thus, we only select 10 non-aggregation complex queries from QALD for evaluation. Two of them are as shown in Table 5 for case study.

All sample queries are shown in Appendix.

6.2 Effectiveness Study

In this section, we compare our method with a classical keyword search algorithm BANKS [2] over DBLP and Yago to show the effectiveness of our method. Furthermore, since each resource in DBPedia is annotated by Wikipedia documents, so we design a stronger baseline named as “Annotated SPARQL” for DBPedia.

6.2.1 Case Study

We show six sample queries in Table 3, 4 and 5 for the case study.

DBLP. Let us consider the two sample queries in DBLP. The top-3 results answered by the SK query and BANKS over DBLP are shown in Table 6.

Q₃: Which researchers on keyword search published papers in VLDB 2004 and DEXA 2005?

The three results returned in SK query are three researchers named “Kesheng Wu”, “Jeffrey Xu Yu” and “Maurice van Keulen”. All of them published paper in both VLDB 2004 and DEXA 2005. As well, they wrote papers about keyword search before. However, the first three results returned by traditional keyword search are “Katsumi Tanaka”, “Mong-Li Lee” and “Reda Alhajj”. Although all

²<http://sw.de.ri.org/aharth/2004/07/dblp/>

³<http://www.mpi-inf.mpg.de/yago-naga/yago/>

⁴<http://downloads.dbpedia.org/3.7/en/>

⁵<http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=challenge&q=2>

	Query Sematic	SK Query	
		SPAQRL	Keywords
Q3	Which researchers on keyword search published papers in VLDB 2004 and DEXA 2005?	Select ?person where { ?paper year 2004; ?paper booktitle VLDB; ?paper1 year 2005; ?paper1 booktitle DEXA; ?paper1 creator ?person; ?paper dc:creator ?person;}	keyword search
Q4	Which papers in KDD 2005 about concept-drifting are written by Jiawei Han?	Select ?paper where { ?paper year 2003; paper booktitle KDD; ?paper creator ?person; ?person name Jiawei Han;}	concept-drifting

Table 3: Sample DBLP Queries for Case Study

	Query Sematic	SK Query	
		SPAQRL	Keywords
Q ₁	Which actors/actresses played in Philadelphia are mostly related to Academy Award and Golden Globe Award?	Select ?p where{ ?p type actor; ?p actedIn ?f; ?f label "Philadelphia"; }	Academy Award, Golden Globe Award
Q ₂	Which Turing Award winners in the field of database are mostly related to Toronto?	Select ?p where { ?p type scientist; ?p hasWonPrize ?a; ?a label "Turing Award";}	Toronto, database

Table 4: Sample Yago Queries for Case Study

of them are interested in keyword search, none of them published paper in VLDB 2004.

Q₄: Which papers in KDD 2005 about concept-drifting are written by Jiawei Han?

The first result returned in SK query is a paper named "Mining concept-drifting data streams using ensemble classifiers". This paper was written by Jiawei Han and published in KDD 2005. This is a paper closely related to concept-drifting. This is the best answer to query Q₂. The other two results of SK queries are still two paper written by Jiawei Han and published in KDD 2005. In contrast, the first two results returned by traditional keyword search are two papers about concept-drifting, but none of them was published in KDD 2005 or written by Jiawei Han. The third result of traditional keyword search is a researcher, which is more unrelated to the query.

Yago. Let us consider the two sample queries in Yago. The top-3 results answered by the SK query and BANKS over Yago are shown in Table 7.

Q₁: Which actors/actresses played in Philadelphia are mostly related to Academy Award and Golden Globe Award?

We have analyzed query Q₁ in Section 1. For comparison, we use keywords {actors, actresses, Philadelphia, Academy Award, Golden Globe Award} for keyword search. Generally, SK query returns more reasonable answers than the traditional keyword search. In contrast, the first two results returned traditional keyword search are "Grace Kelly" and "George Cukor". Grace Kelly lived in Philadelphia, and George Cukor is also an actor that directed the film, *The Philadelphia Story*, in 1940.

Q₂: Which Turing Award winners are mostly related to Toronto?

The first result returned in SK query is "Stephen Cook". As we know, Stephen Cook is a professor in University of Toronto. He won the Turing award for his contributions to complexity theory. This is the best answer to query Q₂. The second answer is "William Kahan". Prof. William Kahan was born in Toronto and won the Turing award for his contributions to the numerical analysis algo-

	Query Sematic	SK Query	
		SPAQRL	Keywords
Q ₃	Which states of Germany are governed by the Social Democratic Party?	Select ?s where { ?s country ?g; ?g name "Germany";}	Social Democratic Party
Q ₄	Which monarchs of the United Kingdom were married to a German?	Select ?u where { ?u spouse ?s; ?s birthPlace ?c; ?c name "Germany";}	United Kingdom, monarch

Table 5: Sample QALD Queries over DBpedia for Case Study

	Top-3 Results of SK Query	Top-3 Results of BANKS
Q3	Kesheng Wu	Katsumi Tanaka
	Jeffrey Xu Yu	Mong-Li Lee
	Maurice van Keulen	Reda Alhajj
Q4	Mining concept-drifting data streams using ensemble classifiers	On Reducing Classifier Granularity in Mining Concept-Drifting Data Streams.
	CLOSET+: searching for the best strategies for mining frequent closed itemsets.	ACE: Adaptive Classifiers-Ensemble System for Concept-Drifting Environments.
	CloseGraph: mining closed frequent graph patterns.	Baile Shi

Table 6: Effectiveness Results for Sample DBLP Queries

rithm. The third one is "Kenneth E. Iverson". Prof. Kenneth E. Iverson also received the Turing Award. He was died in Toronto.

In contrast, the first two results returned by traditional keyword search are "English Language" and "Princeton University". Obviously, they are non-informative results. Here, keywords for keyword search that we use are {Turing Award, winners, Toronto}.

	Top-3 Results of SK Query	Top-3 Results of BANKS
Q ₁	Denzel Washington	Grace Kelly
	Joanne Woodward	George Cukor
	Antonio Banderas	Joanne Woodward
Q ₂	Stephen Cook	English language
	William Kahan	Princeton University
	Kenneth E. Iverson	Turing Award

Table 7: Effectiveness Results for Sample Yago Queries

DBpedia & QALD. Let us consider the two sample QALD queries over DBpedia. The top-3 results answered by BANKS, the SK query and "Annotated SPARQL" over DBpedia are shown in Table 8.

	Top-3 Results of SK Query	Top-3 Results of BANKS	Top-3 Results of Annotated SPARQL
Q ₃	Hanau	Australia	Hans-Ulrich Rudel
	Hanhofen	Bombardier Transportation	Hans Dauser
	Hanover	Canada	Hans Heidtmann
Q ₄	William IV of the United Kingdom	2004 Amsterdam Admirals season	William IV of the United Kingdom
	Carl XVI Gustaf of Sweden	2004 Berlin Thunder season	Beatrix of the Netherlands
	Beatrix of the Netherlands	2004 Cologne Centurions season	Switzerland

Table 8: Effectiveness Results over DBpedia for Sample QALD Queries

Q₃: Which states of Germany are governed by the Social Democratic Party?

The first three results returned in SK query are three places in Germany and governed by the Social Democratic Party. However, the first three results returned in annotated SPARQL are three members of the Social Democratic Party in Germany. The first three re-

		NDCG@3	NDCG@5	NDCG@10
Yago	BANKS	0.3455	0.39	0.4643
	SK query	0.815	0.868	0.872
DBLP	BANKS	0.7143	0.684	0.685
	SK query	0.93	0.8867	0.8738

Table 9: Average NDCG Values

sults returned by traditional keyword search are three place far from Germany. Hence, the results of SK queries are more informative than the other two methods. Here, keywords for keyword search are {*state, Germany, govern, SocialDemocraticParty*}, which are given in QALD.

Q₄: Which monarchs of the United Kingdom were married to a German?

The first result of both SK query and annotated SPARQL are William IV of the United Kingdom, which is the best answer. The other two results of SK query are still two royals in Europe. However, the third result of annotated SPARQL is a European nation. In addition, the first three results returned by traditional keyword search are non-informative results. Here, keywords for keyword search are {*United Kingdom, monarch, married, German*}, which are also given in QALD.

6.2.2 NDCG@k over Yago and DBLP

In order to quantify the effectiveness of SK query, we evaluate the NDCG (Normalized Discounted Cumulative Gain [16]) of both SK query and the keyword search. Since there are no golden standards, we invite 10 volunteers to judge the result quality. Specifically, we ask each volunteer to rate the goodness of the results returned by SK query and the keyword search method. The score is between 1 and 5. Higher the score, better the result.

Table 9 reports NDCG@k values by varying k from 3 to 10 in both Yago and DBLP. SK query outperforms the traditional keyword search by 20%-50%. Furthermore, we find that the gap in Yago is larger than that in DBLP. The reason is that Yago has more complex schema than DBLP. Thus, keywords may result in more ambiguity in Yago than in DBLP. It means that the superiority of SK query is more pronounced in semantic-rich data.

6.2.3 MAP over DBPedia

Since QALD provides the standard answers of each queries, we evaluate the MAP (Mean Average Precision [30]) to compare the SK query with BANKS and “Annotated SPARQL”.

	BANKS	Annotated SPARQL	SK query
MAP	0.012	0.192	0.205

Table 10: MAP Value over DBPedia & QALD

Table 10 reports MAP values of our ten QALD queries. Both SK query and annotated SPARQL outperforms the traditional keyword search by a order of magnitude. The MAP value of the “annotated SPARQL” is smaller than the SK query. This is because that the “annotated SPARQL” can do well when the documents associated with the matches contains the keywords. In other words, the “annotated SPARQL” can work, only when the relation between the matches and the keywords is explicit. However, in practice, the relation between the matches and the keywords is often implicit. Then, the SK query do better.

6.3 Efficiency Study

In this section, we evaluate the efficiency of SK query in large real graphs. Here, the default number of returned results is set to be 10.

6.3.1 Offline Performance

We report the index size and index construction time in Table 14. Since our structural index is based on the efficient sequential pattern mining, we can finish the structural index construction in several minutes.

	Index Construction Time(s)	Index Size(MB)
DBLP	77.885	377.977
Yago	176.67	844.066
DBPedia	600.263	283.559

Table 14: Index Size and Index Construction time

6.3.2 Pruning Effect of Structural Index

Based on the indices introduced in Section 4, we can avoid many times to call Algorithm 2 for graph matching by pruning many unsatisfied vertices. Moreover, the vertices which are too far to be in a final answer can be safely pruned. In this experiment, we report the pruning efficiency of our structural index. We make a comparison of the number of graph matching operations that the advanced backward search accessed and the number of graph matching operations that the naive backward search accessed.

Tables 11, 12 and 13 show the number of graph matching operations on DBLP, Yago and DBPedia. The number of graph matching operations in advanced backward search is not less than the basic backward search. In most case, we avoid a large number number of graph matching operations.

6.3.3 Online Performance

In this section, we evaluate the efficiency of our method. Figure 9 shows the time cost of the three methods.

As shown in Figure 9, our method outperforms the baseline method by 2 or more times in most case. Especially for Q_3 on Yago and Q_2 , Q_5 on DBLP, our method only takes a fifth of the exhaustive-computing. This is because that the matches of these SPARQLs are close to vertices containing keywords. Thus, the query processing can terminate soon.

Note that, because our inverted index for keywords are stored in disk, keywords mapping will cost much time and takes up a large part of the total time. Hence, it is difficult for our method to improve the efficiency too much.

7. RELATED WORK

For SPARQL query, there have been many works to study it, such as [35, 1, 22, 23, 41, 5, 38]. Some of them [1, 5] store the RDF triples into RDBMS and answer the SPARQL via join operations. RDF-3x [22, 23] and Hexastore [35] create indexes for each permutation of subject, predicate and object. Since an RDF dataset can also be modeled as a graph, Trinity.RDF [38] and gStore [41] deem answer the SPARQL in an RDF dataset as finding the subgraph matches over an RDF graph. Trinity.RDF and gStore design a subgraph match algorithm similar to VF2 [8] to answer the SPARQL query. VF2 [8] is an early efforts for subgraph isomorphism check. VF2 starts with a vertex and explore to a vertex connected from the already matched query vertices one by one.

For keyword search, existing keyword search techniques over RDF graphs can be classified into the following two categories. The first kind of methods [34, 29, 12, 13] interpret keywords as SPARQL queries, and then retrieve results by involving existing SPARQL query engines. Another kind of methods aim to find the small-size substructures (in RDF graphs) that contains all keywords. The top-k substructures, such as trees [2, 17, 15, 10, 19, 20, 32], cliques [18], computed on the basis of a scoring function are returned to users.

	Q_1	Q_2	Q_3	Q_4	Q_5
Naive Backward Search	292268	21885	254674	872426	2747
Our Approach	354	5	1684	669	1548

Table 11: The Number of Graph Matching Operations on DBLP

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
Naive Backward Search	600283	563736	301	167958	231210	271929	94012	254848
Our Approach	6	55	269	5414	32	9	292	15

Table 12: The Number of Graph Matching Operations on Yago

There are also many approaches to mine some frequent patterns to build indices in graph database [37, 36, 39]. Among these works, gIndex [37] and gSpan [36] can be applied to small graphs in a database of multiple graphs, but not support mining patterns in a single graph. GADDI [39] tries to finding all the matches of a query graph in a given large graph, but it can only support a graph with thousands of vertices while recent RDF data graph may have hundred thousands of entities.

To the best of our knowledge, although there exist a few previous works [11, 33] for the hybrid query combined SPARQL and keywords, there has been no existing work on SK query defined as the above. Elbassuoni et al. [11] assumes that each RDF triple may have associated text passages. Then, Elbassuoni et al. extend the triple patterns in SPARQL with keyword conditions. Moreover, CE^2 [33] assumes that each resource associate with a document. Then, CE^2 extend the variables in SPARQL with keyword conditions. Nonetheless, most current RDF datasets do not provide neither text passages to annotate triples nor documents to annotate resources. In summary, both of these methods cannot handle our example queries. Also, the SK query that we define can apply to most existing RDF datasets.

As well, in [26], the authors define a new query language that blends keyword search with structured query processing. [28] utilizes some given kinds of SPARQL to improve the result of object retrieval. Moreover, [3, 4] try to extend keyword search with semantics. Zou et al. [40] translate natural language questions into SPARQL queries.

8. CONCLUSIONS

In this paper, we have proposed a new kind of query (SK query) that integrates SPARQL and keywords. To handle this kind of query, we firstly introduce a basic method based on the backward search. However, this basic solution faces several performance issues. Hence, we build up a structural index. Our structural index is based on frequent star pattern in RDF data. By using the indices, we propose an advanced strategy to deal with SK queries. Finally, with three real RDF datasets, we demonstrate that the our method can outperform the baseline both in effectiveness and efficiency.

9. REFERENCES

- [1] D. J. Abadi, A. Marcus, S. Madden, and K. Hollenbach. Sw-store: a vertically partitioned dbms for semantic web data management. *VLDB J.*, 18(2):385–406, 2009.
- [2] B. Aditya, G. Bhalotia, S. Chakrabarti, A. Hulgeri, C. Nakhe, Parag, and S. Sudarshan. Banks: Browsing and keyword searching in relational databases. In *VLDB*, pages 1083–1086, 2002.
- [3] R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli. Hybrid search: Effectively combining keywords and semantic searches. In *ESWC*, pages 554–568, 2008.
- [4] N. Bikakis, G. Giannopoulos, T. Dalamagas, and T. K. Sellis. Integrating keywords and semantics on document annotation and search. In *OTM Conferences (2)*, pages 921–938, 2010.
- [5] M. A. Bornea, J. Dolby, A. Kementsietsidis, K. Srinivas, P. Dantressangle, O. Udrea, and B. Bhattacharjee. Building an efficient rdf store over a relational database. In *SIGMOD Conference*, pages 121–132, 2013.
- [6] R. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 19(3):370–383, 2007.
- [7] R. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 19(3):370–383, 2007.
- [8] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(10):1367–1372, 2004.
- [9] D. Deng, G. Li, and J. Feng. A pivotal prefix based filtering algorithm for string similarity search. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22–27, 2014*, pages 673–684, 2014.
- [10] B. Ding, J. X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin. Finding top-k min-cost connected trees in databases. In *ICDE*, pages 836–845, 2007.
- [11] S. Elbassuoni, M. Ramanath, R. Schenkel, and G. Weikum. Searching rdf graphs with sparql and keywords. *IEEE Data Eng. Bull.*, 33(1):16–24, 2010.
- [12] H. Fu and K. Anyanwu. Effectively interpreting keyword queries on rdf databases with a rear view. In *International Semantic Web Conference (I)*, pages 193–208, 2011.
- [13] H. Fu, S. Gao, and K. Anyanwu. Codi: context-sensitive keyword query interpretation on rdf databases. In *WWW (Companion Volume)*, pages 209–212, 2011.
- [14] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, San Francisco, 1979.
- [15] H. He, H. Wang, J. Yang, and P. S. Yu. Blinks: ranked keyword searches on graphs. In *SIGMOD Conference*, pages 305–316, 2007.
- [16] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*, pages 41–48, 2000.
- [17] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar. Bidirectional expansion for keyword search on graph databases. In *VLDB*, pages 505–516, 2005.
- [18] M. Kargar and A. An. Keyword search in graphs: Finding r-cliques. *PVLDB*, 4(10):681–692, 2011.
- [19] G. Kasneci, M. Ramanath, M. Sozio, F. M. Suchanek, and G. Weikum. Star: Steiner-tree approximation in relationship graphs. In *ICDE*, pages 868–879, 2009.
- [20] W. Le, F. Li, A. Kementsietsidis, and S. Duan. Scalable keyword search on large RDF data. *IEEE Trans. Knowl. Data Eng.*, 26(11):2774–2788, 2014.
- [21] M. Ley and P. Reuther. Maintaining an online bibliographical database: The problem of data quality. In *EGC*, pages 5–10, 2006.
- [22] T. Neumann and G. Weikum. Rdf-3x: a risc-style engine for rdf. *PVLDB*, 1(1):647–659, 2008.
- [23] T. Neumann and G. Weikum. x-rdf-3x: Fast querying, high update rates, and consistency for rdf databases. *PVLDB*, 3(1):256–263, 2010.
- [24] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *ICDE*, pages 215–224, 2001.
- [25] J. Pérez, M. Arenas, and C. Gutiérrez. Semantics and complexity of SPARQL. *ACM Trans. Database Syst.*, 34(3), 2009.
- [26] J. Pound, I. F. Ilyas, and G. E. Weddell. Expressive and flexible

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
Naive Backward Search	31083	136777	19904	847	19454	40302	5076	23422	16079	2786
Our Approach	13824	3941	4769	847	40	89	18	23422	16079	1

Table 13: The Number of Graph Matching Operations on DBPedia

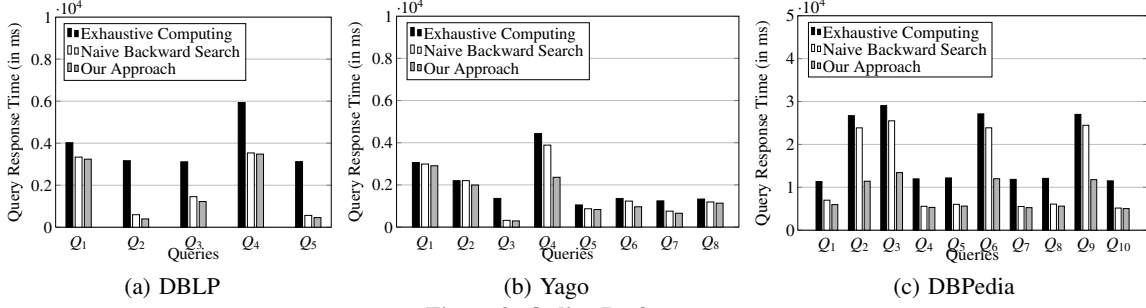


Figure 9: Online Performance

access to web-extracted data: a keyword-based structured query language. In *SIGMOD Conference*, pages 423–434, 2010.

- [27] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.
- [28] A. Tonon, G. Demartini, and P. Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *SIGIR*, pages 125–134, 2012.
- [29] T. Tran, H. Wang, S. Rudolph, and P. Cimiano. Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data. In *ICDE*, pages 405–416, 2009.
- [30] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR*, pages 11–18, 2006.
- [31] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.
- [32] D. Wang, L. Zou, W. Pan, and D. Zhao. Keyword graph: Answering keyword search over large graphs. In *Advanced Data Mining and Applications, 8th International Conference, ADMA 2012, Nanjing, China, December 15-18, 2012. Proceedings*, pages 635–649, 2012.
- [33] H. Wang, T. Tran, C. Liu, and L. Fu. Lightweight integration of ir and db for scalable hybrid search with integrated ranking support. *J. Web Sem.*, 9(4):490–503, 2011.
- [34] H. Wang, K. Zhang, Q. Liu, T. Tran, and Y. Yu. Q2semantic: A lightweight keyword interface to semantic search. In *ESWC*, pages 584–598, 2008.
- [35] C. Weiss, P. Karras, and A. Bernstein. Hexastore: sextuple indexing for semantic web data management. *PVLDB*, 1(1):1008–1019, 2008.
- [36] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *ICDM*, pages 721–724, 2002.
- [37] X. Yan, P. S. Yu, and J. Han. Graph indexing: A frequent structure-based approach. In *SIGMOD Conference*, pages 335–346, 2004.
- [38] K. Zeng, J. Yang, H. Wang, B. Shao, and Z. Wang. A distributed graph engine for web scale rdf data. *PVLDB*, 6(4):265–276, 2013.
- [39] S. Zhang, S. Li, and J. Yang. Gaddi: distance index based subgraph matching in biological networks. In *EDBT*, pages 192–203, 2009.
- [40] L. Zou, R. Huang, H. Wang, J. X. Yu, W. He, and D. Zhao. Natural language question answering over RDF: a graph data driven approach. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 313–324, 2014.
- [41] L. Zou, J. Mo, L. Chen, M. T. Özsu, and D. Zhao. gstore: Answering sparql queries via subgraph matching. *PVLDB*, 4(8):482–493, 2011.

aggregation QALD queries over DBPedia to evaluate our method. All QALD queries over DBPedia are shown in Table 16.

APPENDIX

A. QUERIES IN EXPERIMENTS

Table 15 shows all of our sample queries over Yago and DBLP. Here, since our institution is in China, most volunteers that we invite are Chinese. Hence, some sample queries are about China.

For more reasonable experiments, so we also sample 10 non-

		Query Sematic	SK Query	
			SPAQRL	Keywords
DBLP	Q1	Which researchers on keyword search published papers in VLDB 2004 and DEXA 2005?	Select ?person where { ?paper year 2004 ?paper booktitle VLDB ?paper1 year 2005 ?paper1 booktitle DEXA ?paper1 creator ?person ?paper creator ?person}	keyword search
	Q2	Which papers in KDD 2005 about concept-drifting are written by Jiawei Han?	Select ?paper where { ?paper year 2003 paper booktitle KDD ?paper creator ?person ?person name Jiawei Han}	concept-drifting
	Q3	Who wrote a paper in ICDM 2005 with others and knew Tamer?	Select ?person1 where { ?paper year 2005; ?paper booktitle "ICDM"; ?paper creator ?person1; ?paper dc:creator ?person2}	Tamer
	Q4	Who wrote a paper VLDB 2005 and kept a good relationship to Jian Pei and Wen Jin?	Select ?person2 where { ?paper year "2005"; ?paper booktitle "VLDB"; ?paper creator ?person2;}	Jian Pei, Wen Jin
	Q5	Which two researchers did research about skyline and coauthored a paper in VLDB 2005?	Select ?person1, ?person2 where { ?paper year "2005"; ?paper booktitle "VLDB"; ?paper dc:creator ?person1; ?paper dc:creator ?person2}	Skyline
Yago	Q1	Which actresses played in Philadelphia are mostly related to Academy Award and Golden Globe Award?	Select ?p where{ ?p type actor; ?p actedIn ?f; ?f label "Philadelphia"; }	Academy Award, Golden Globe Award
	Q2	Which Turing Award winners in the field of database are mostly related to Toronto?	Select ?p where { ?p type scientist; ?p hasWonPrize ?a; ?a label "Turing Award";}	Toronto, database
	Q3	Which Microsoft's products are about SDK?	Select ?c where { ?c type company; ?c label "Microosft"; ?c created ?s;}	SDK
	Q4	Which English film producers did act in a Comedy film and relate to Peking University?	Select ?p where { ?p actedIn ?f1; ?p y:created ?f2; ?f1 type ComedyFilms; ?f2 y:hasProductionLanguage English;}	Peking University
	Q5	Which top members of Communist Party of China are related Kissinger?	Select ?p where { ?p isAffiliatedTo ?u; ?u label "Communist Party of China"; ?p type Politician;}	Kissinger
	Q6	Whose father was United States Army generals and took part in Normandy Invasion?	Select ?p1 where { ?p hasChild ?p1; ?p type UnitedStatesArmyGenerals ;}	Normandy Invasion
	Q7	Which state generated a Los Angeles Lakers player that relate to Eagle, Colorado?	Select ?p where { ?p bornIn ?c; ?c locatedIn ?s; ?p type LosAngelesLakersPlayers;}	Eagle Colorado
	Q8	Which participants of People's National Congress did graduate from universities in Beijing?	Select ?p where { ?p graduatedFrom ?u; ?u type UniversitiesInBeijing;}	People's National Congress

Table 15: Sample Queries over Yago and DBLP

		Query Sematic	SK Query	
			SPAQRL	Keywords
DBPedia &QALD	Q1	Which states of Germany are governed by the Social Democratic Party?	Select ?s where { ?s country ?g; ?g name "Germany";}	Social Democratic Party
	Q2	Which monarchs of the United Kingdom were married to a German?	Select ?u where { ?u spouse ?s; ?s birthPlace ?c; ?c name "Germany";}	United Kingdom, monarch
	Q3	Which capitals in Europe were host cities of the summer olympic games?	Select ?u where { ?s type Country; ?s capital ?u;}	Olympic games, Europe
	Q4	Who produced films starring Natalie Portman?	Select ?p where { ?f type Film; ?f producer ?p;}	Natalie Portman
	Q5	In which films did Julia Roberts as well as Richard Gere play?	Select ?f where { ?f type Film; ?f starring ?p; ?p name "Roberts, Julia"}	Richard Gere
	Q6	List all episodes of the first season of the HBO television series The Sopranos!	Select ?u where { ?s name "The Sopranos"; ?u series ?s;}	HBO, first
	Q7	In which films directed by Garry Marshall was Julia Roberts starring?	Select ?f where { ?f type Film; ?f director ?p; ?p name "Marshall, Garry";}	Julia Roberts
	Q8	Which software has been developed by organizations founded in California?	Select ?u where { ?c type Organisation; ?u developer ?c; ?u type Software;}	California
	Q9	Which U.S. states possess gold minerals?	Select ?s where { ?s type Place; ?s country ?g; ?g name "the United States";}	gold, mineral
	Q10	Which countries in the European Union adopted the Euro?	Select ?u where { ?u type Country; ?u ethnicGroup European Union;}	Euro

Table 16: Sample QALD Queries over DBPedia